Identification of Bacteria Using Phylogenetic Relationships Revealed by MS/MS Sequencing of Tryptic Peptides Derived from Cellular Proteins

Jacek P. Dworzanski

Geo-Centers, Inc.,
Aberdeen Proving Ground, MD 21010-0068

Samir Deshpande[1]; Rui Chen[2]; A. Peter Snyder[3];
Liang Li[2] and Charles H. Wick[3]

[1]Science and Technology Corporation, Edgewood, MD 21040; [2]Department of Chemistry, University of Alberta, Edmonton, Alberta T6G 2G2; [3] U.S. Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD 21010-5424

2004 Joint Service Scientific Conference on Chemical & Biological Defense Research
15-17 November 2004; Hunt Valley, Maryland

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
17 NOV 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Identification of Bacteria Using Phylogenetic Relationships Revealed by MS/MS Sequencing of Tryptic Peptides Derived from Cellular Proteins	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Science and Technology Corporation, Edgewood, MD 21040**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001849, 2004 Scientific Conference on Chemical and Biological Defense Research. Held in Hunt Valley, Maryland on 15-17 November 2004 . , The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**27**	

National Institute of Allergy and Infectious Diseases
National Institute of Health

Category A, B & C Priority Bacterial Pathogens

Category A

1. *Clostridium botulinum*
2. *Bacillus anthracis* (anthrax)
3. *Francisella tularensis* (tularemia)
4. *Yersinia pestis*

Category B

1. *Brucella species* (brucellosis)
2. *Burkhoderia pseudomallei*
3. *Burkholderia mallei* (glanders)
4. *Campylobacter jejuni*
5. *Clostridium perfringens* (epsilon toxin)
6. *Coxiella burnetti* (Q fever)
7. *Escherichia coli* (diarrheagenic)
8. *Listeria monocytogenes*
9. *Rickettsia prowazekii* (typhus fever)
10. *Salmonella*
11. *Shigella* species
12. *Staphylococcus aureus* (enterotoxin B)
13. *Vibrios* (pathogenic)
14. *Yersinia enterocolitica*

Category C

1. *Mycobacterium tuberculosis* (multiple drug resistant)
2. *Rickettsias* (other)

Genomes of all above organisms have been sequenced

5/12/2004 versus {11/12/2004}

Number of Fully Sequenced Genomes of Eubacteria: **145 {178} (From 1995)**

Number of Fully Sequenced Genomes During Last 12 Months: **55 {68}**

Prokaryotic Ongoing Genome Projects: 489 {528}

Archaeal: 28 {27} **Bacterial: 461 {499}**

Example of Ongoing Genome Projects

Bacillus anthracis A1055 (Group C)	TIGR/NIAID
Bacillus anthracis Ames Ancestor	TIGR/NIAID / NSF
Bacillus anthracis Ames Florida	TIGR/NSF
Bacillus anthracis Austrailia 94 (GT55, Group A3a)	TIGR/NIAID
Bacillus anthracis CNEVA-9066 (GT79 Group B2)	TIGR/NIAID
Bacillus anthracis Kruger B (GT87 Group B1)	TIGR/NIAID
Bacillus anthracis Vollum (GT77 Group A4)	TIGR/NIAID
Bacillus anthracis Western N. America (GT3 Group A1a)	TIGR/NIAID
Bacillus anthracis STN	*DOE/JGI*
Bacillus anthracis ZK	*DOE/JGI*

First, some terminology…

Taxonomy - **the science of naming and classifying organisms;**

Classification - **placement of an organism within a scheme relating different groups of organisms;**

Identification - **the determination of whether an organism should be placed within a group of organisms known to fit within some classification scheme;**
(the practical use of classification criteria)

Phylogenetics - **focuses on evolutionary relationships between organisms or genes/proteins**

Phylogenetic Approach

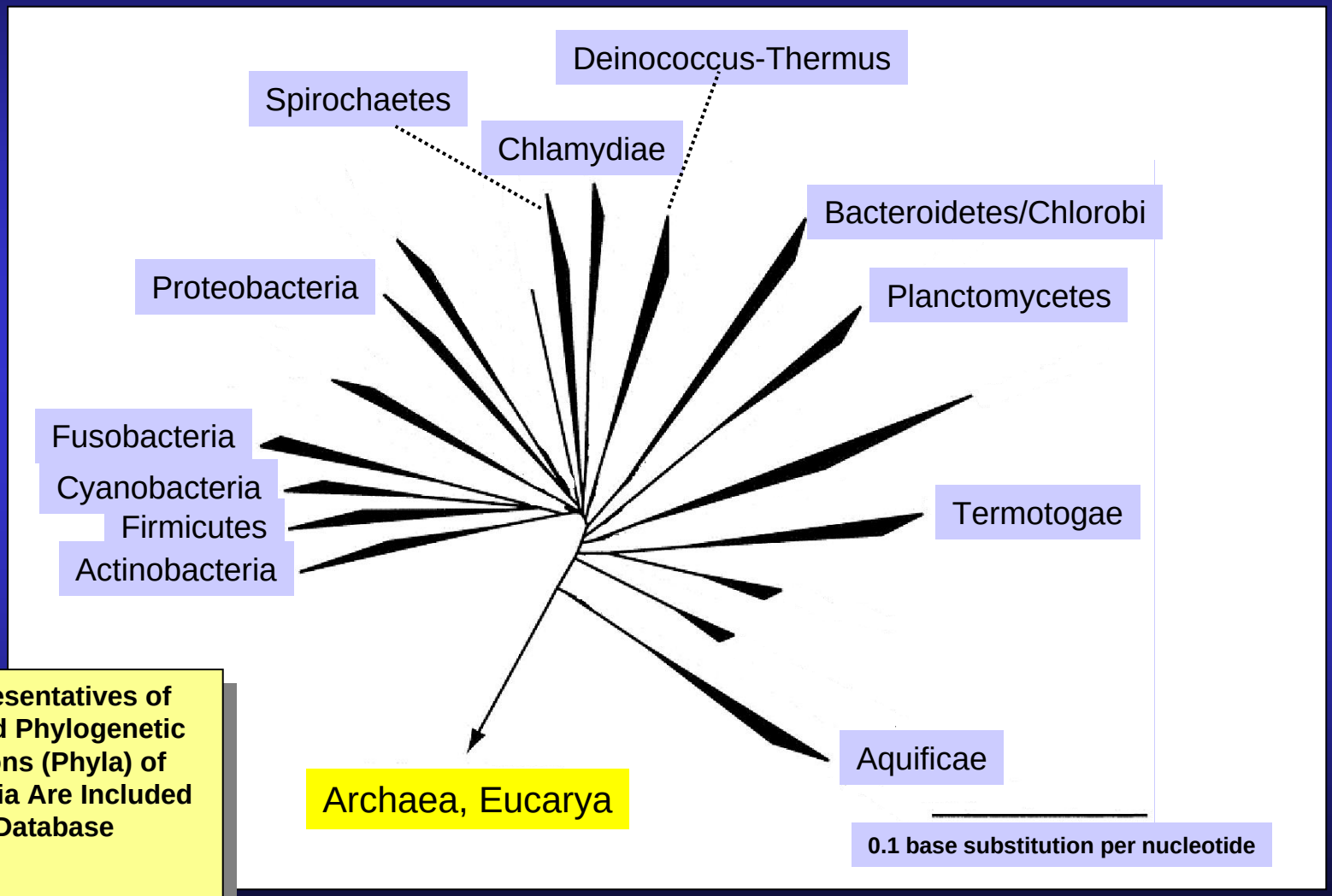
The ideal means of identifying and classifying bacteria would be to compare each gene sequence in a given strain with the gene sequences for every known species.

Taxonomy of Bacteria
(Linnaean System)

Examples: ***Bacillus subtilis* & *Escherichia coli***

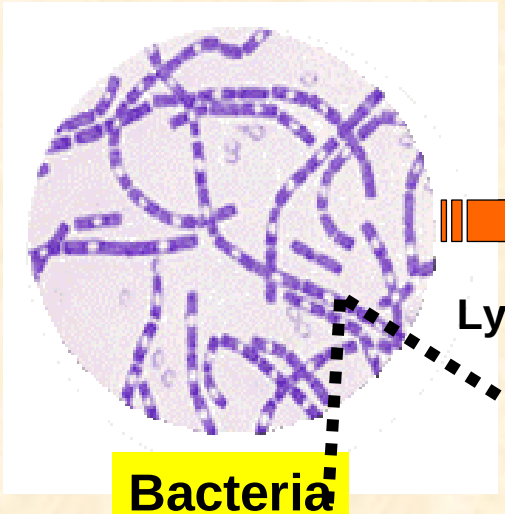
Kingdom	Bacteria	Bacteria
Phylum	Firmicutes	Proteobacteria
Class	Bacilli	γ-Proteobacteria
Order	Bacillales	Enterobacteriales
Family	Bacillaceae	Enterobacteriaceae
Genus	Bacillus	Escherichia
Species	*Bacillus subtilis*	*Escherichia coli*

Universal Phylogenetic Tree of Bacteria Based on SSU rRNA Sequences



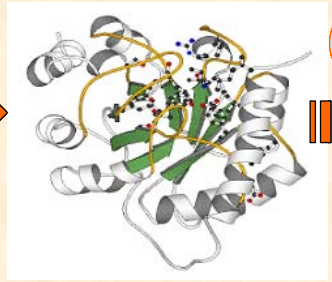
Reepresentatives of Marked Phylogenetic Divisions (Phyla) of Bacteria Are Included in the Database

Approach: (1) Bacterial Sample Processing



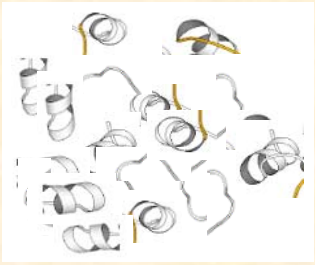
Bacteria

Lysis



Proteins

Digestion
(trypsin)



Peptides

Flow of Genetic Information

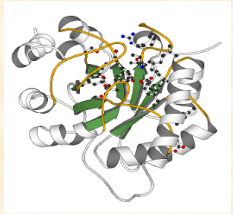
DNA

transcription

mRNA

translation

Protein



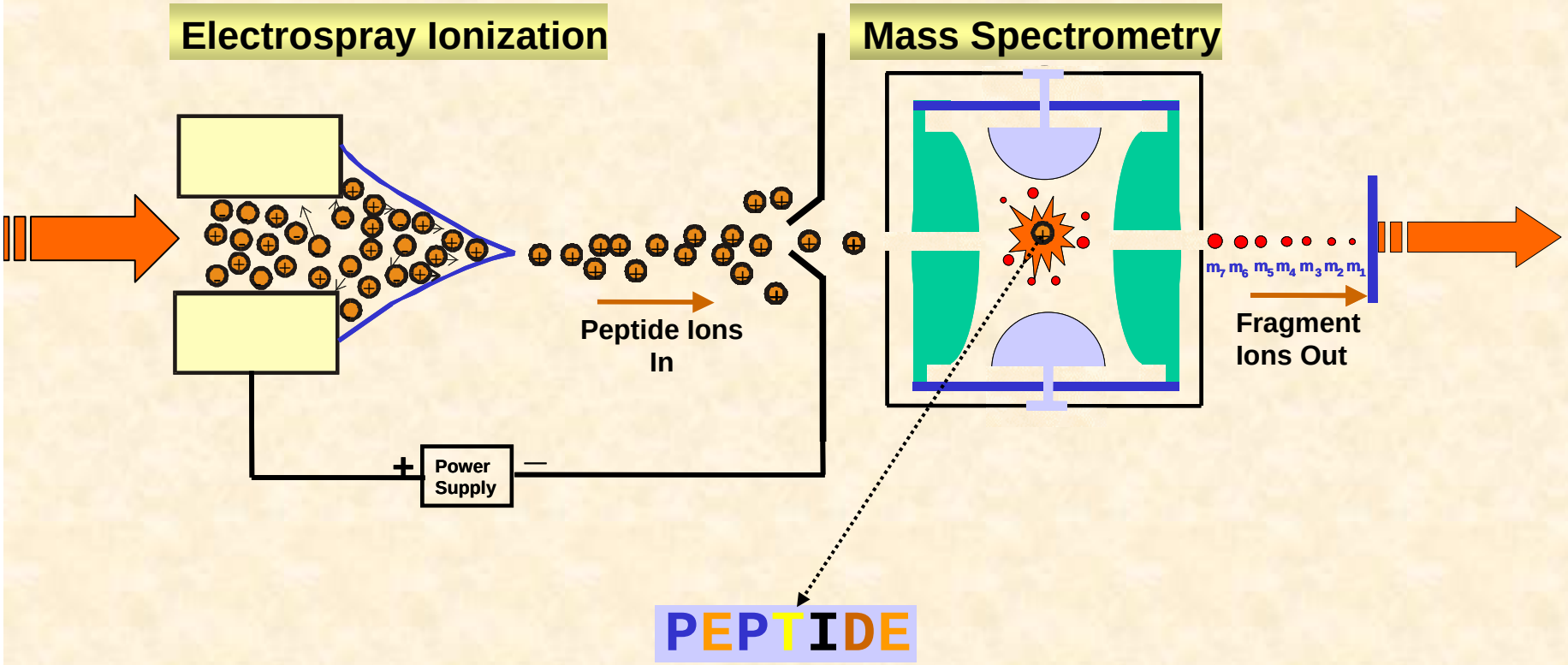
CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

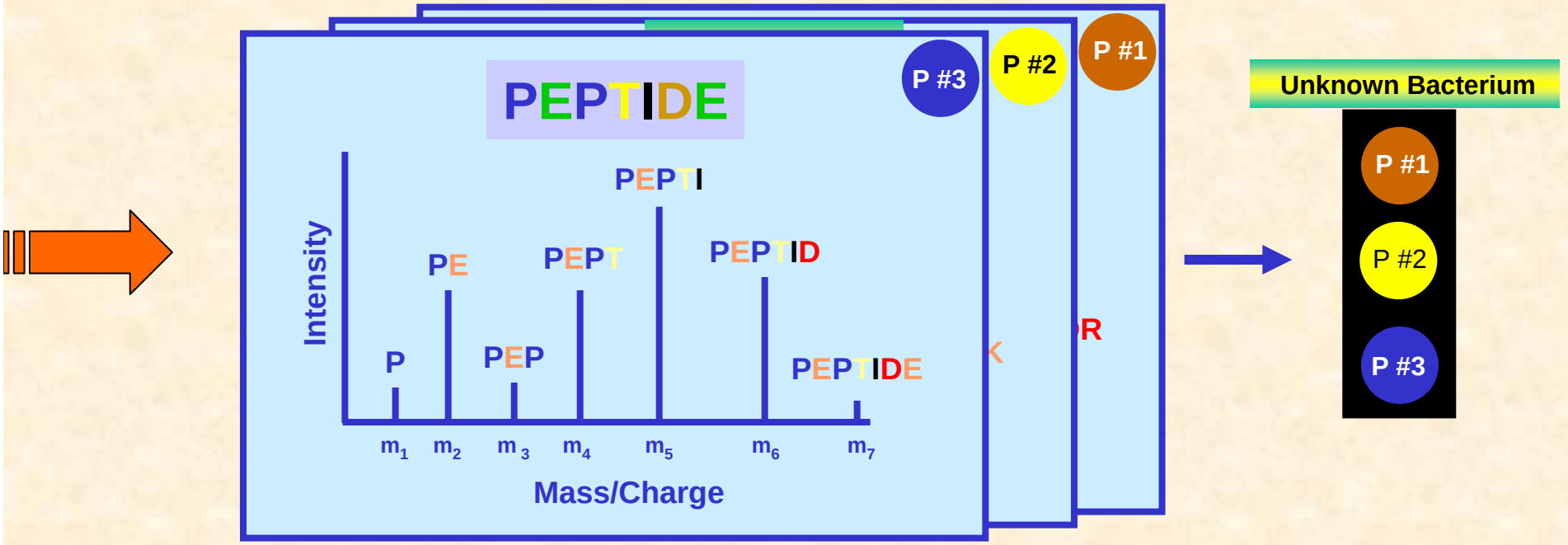
The relation between the sequence of bases in DNA and the sequence of amino acids in a protein

Approach: (2) Tandem Mass Spectrometry of Peptide Ions

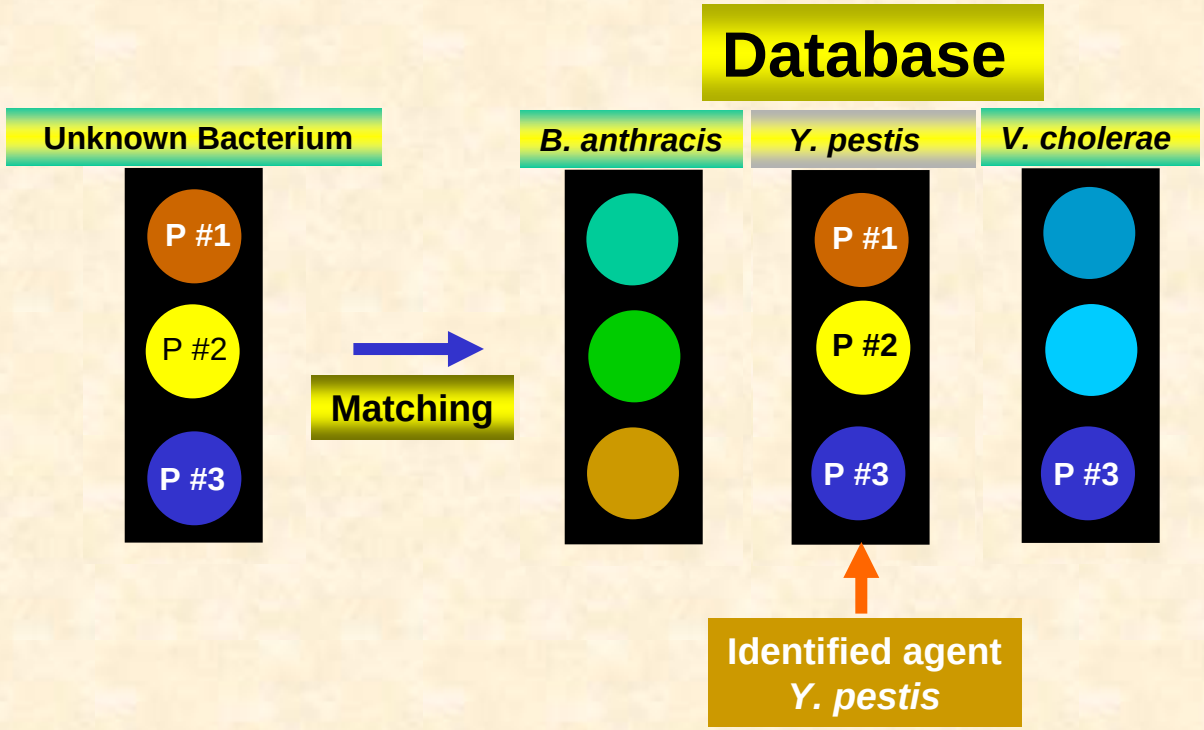


Approach: (3) Sequencing of Peptide Ions

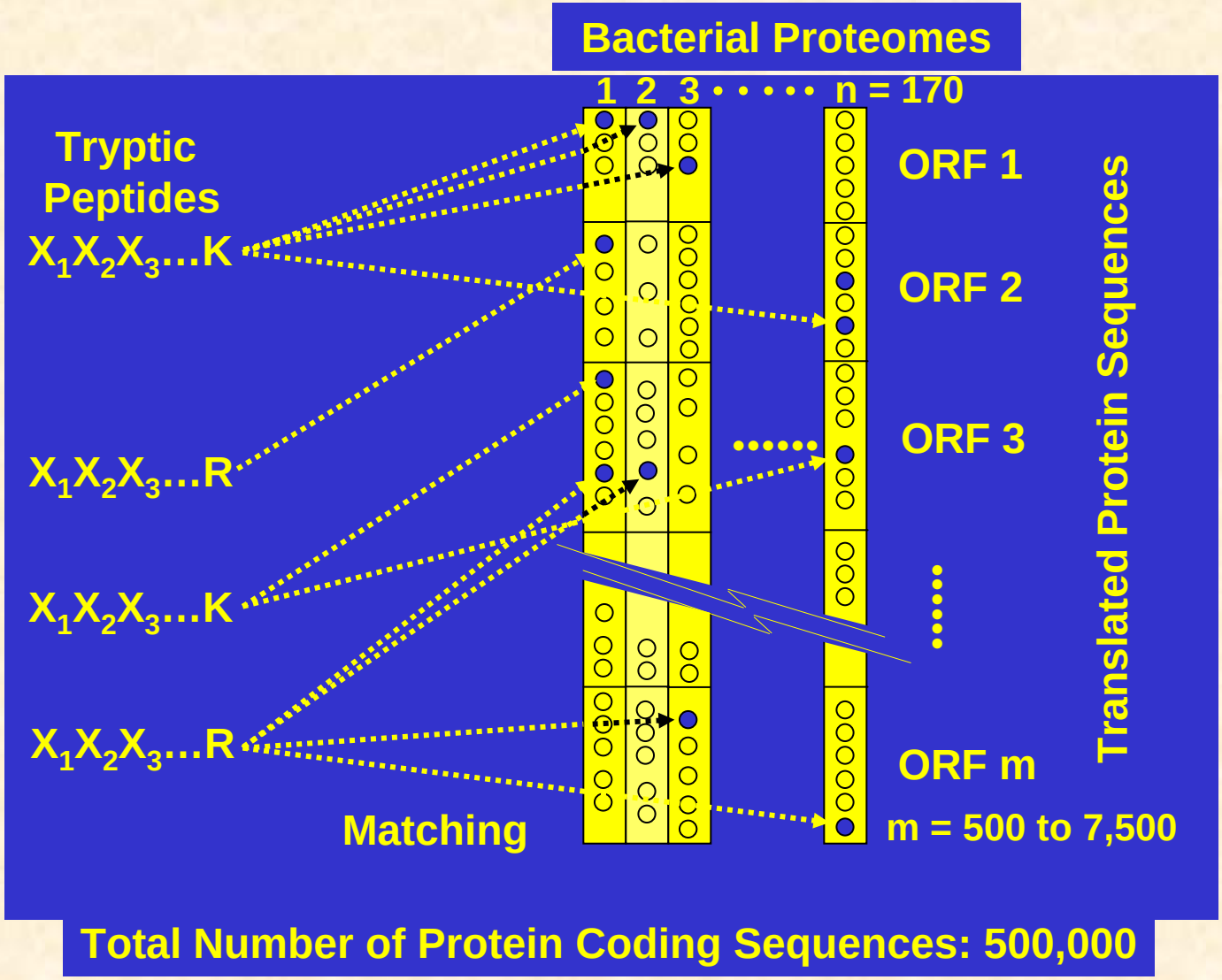
MS/MS Spectra of Peptide Ions



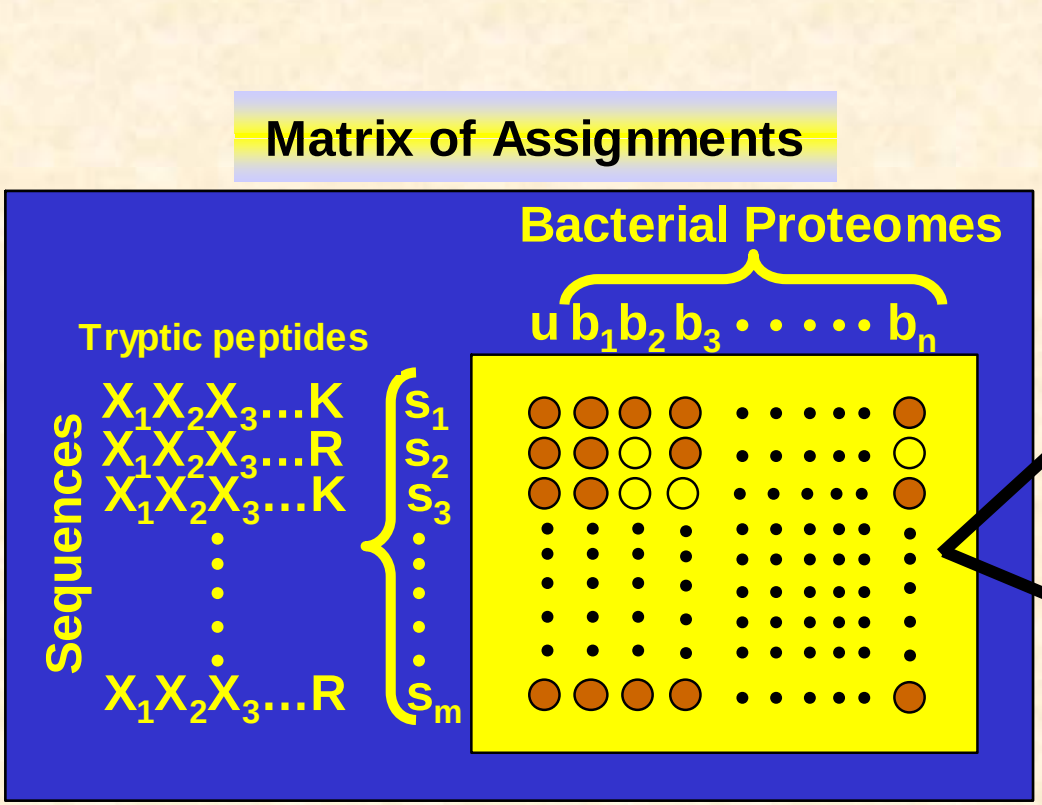
Approach: (4) Matching of Identified Tryptic Peptides to Theoretical Peptides of Database Bacterial Proteomes

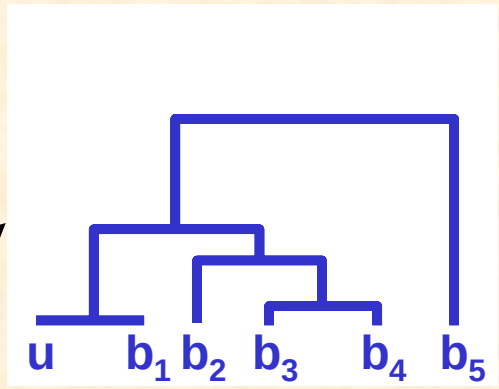


Virtual Array of Peptide Sequences

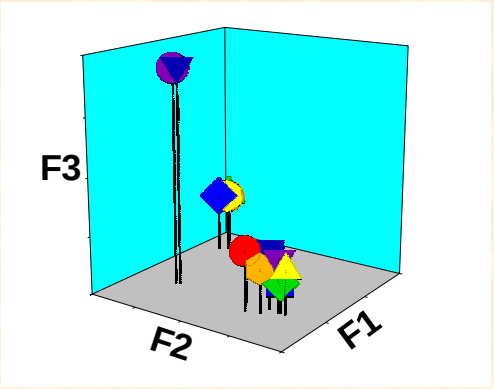


Classification and Identification of Unknown Bacterium

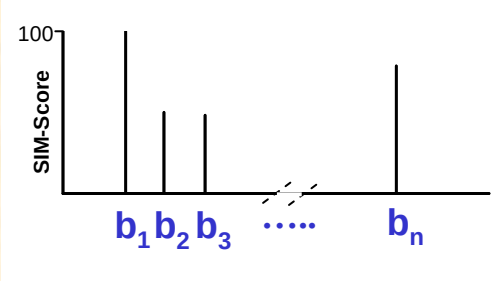




Cluster Analysis



Principal Component Analysis



Affinity Histogram

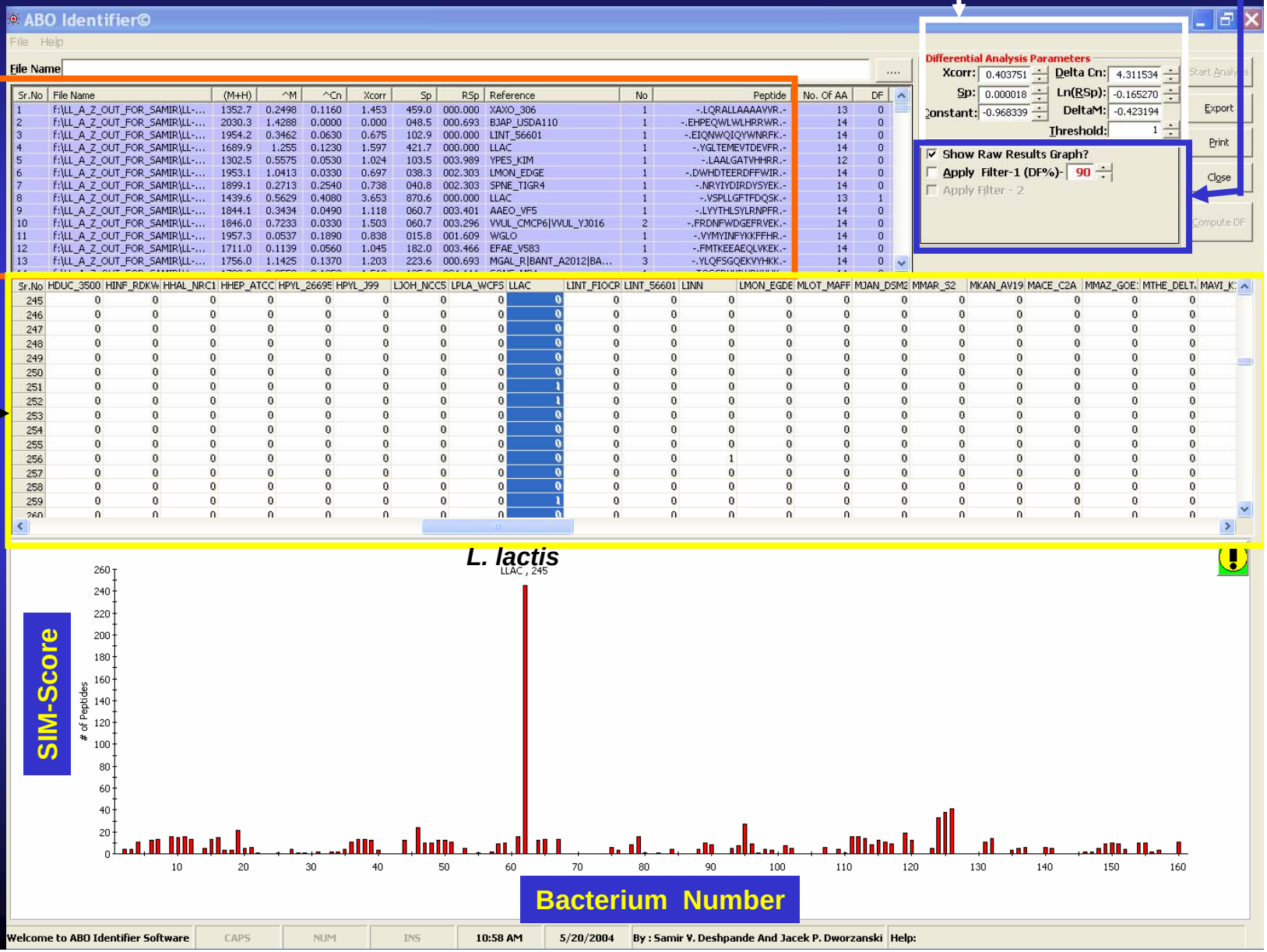
Screen Capture Image Displaying Raw LC-MS/MS Data

SEQUEST Output Data

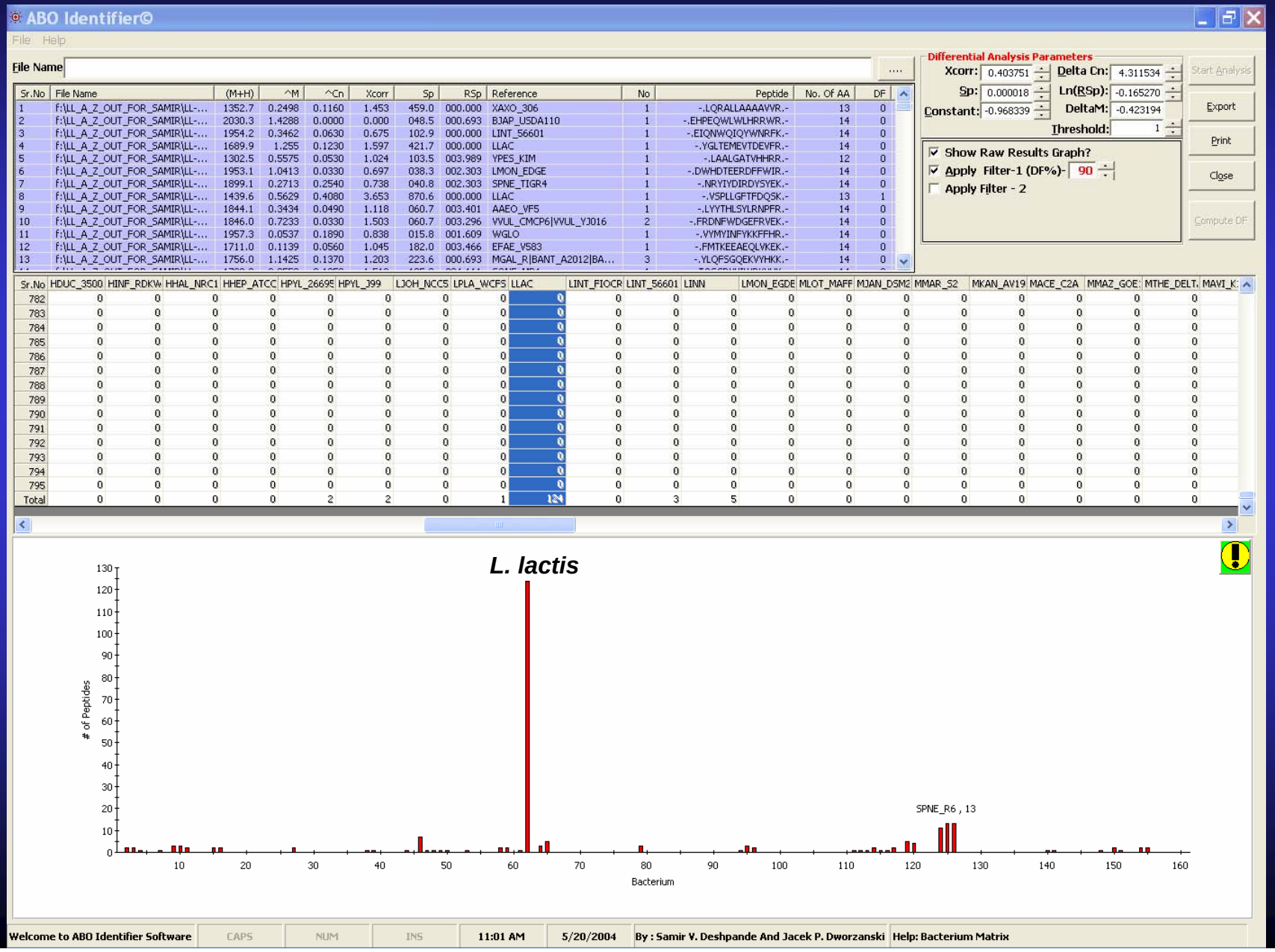
Discriminant Analysis Parameters

Filters

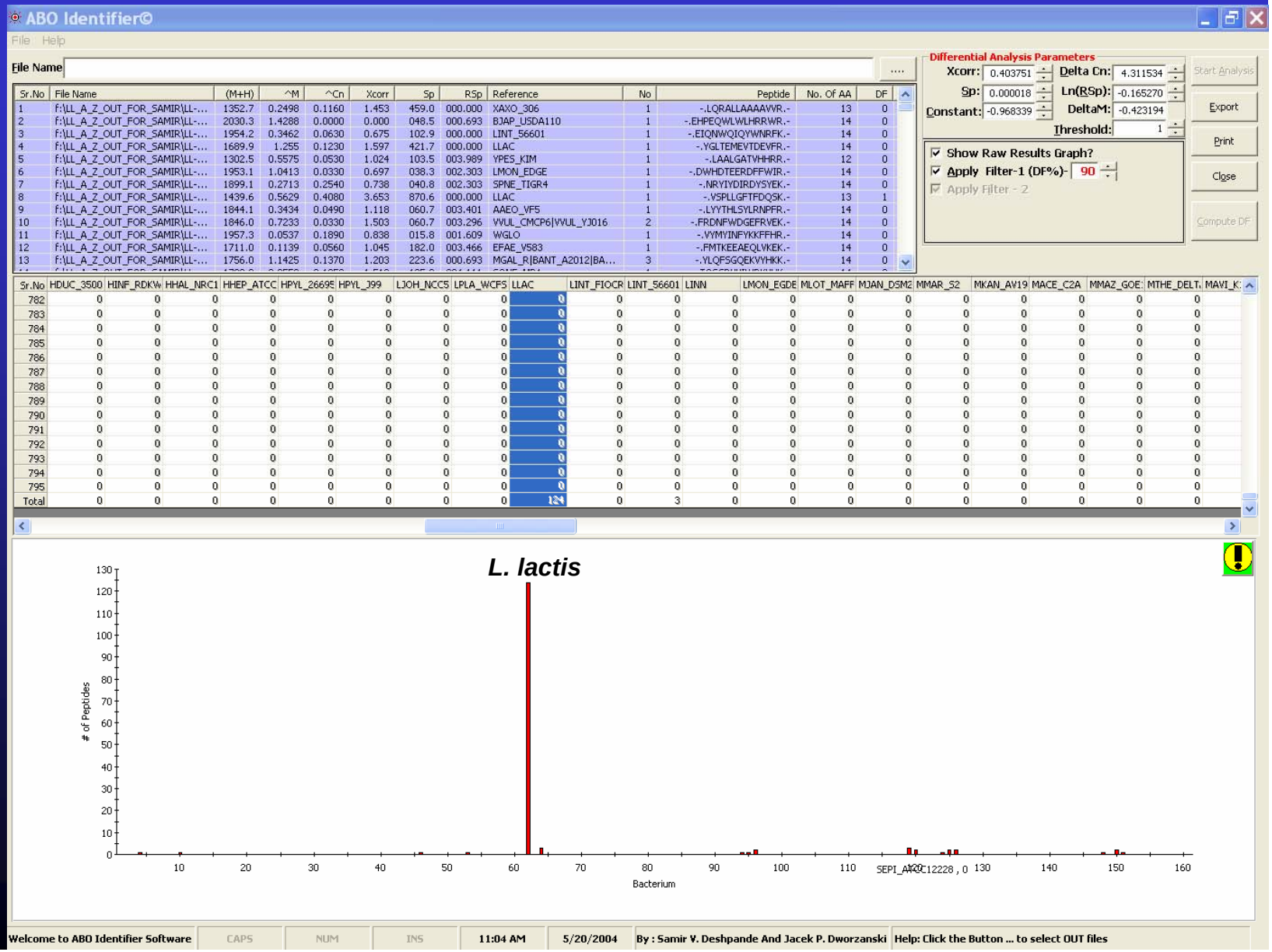
Assignment Matrix



Screen Capture Image Displaying Accepted Peptide Assignments (P = 90%)

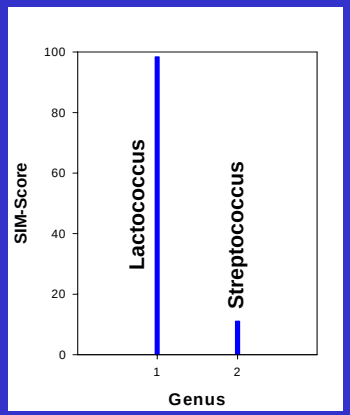


Screen Capture Image Displaying Accepted Peptide Assignments After Removal of Degenerate Peptide Sequences

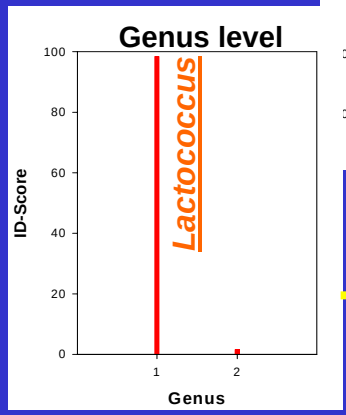


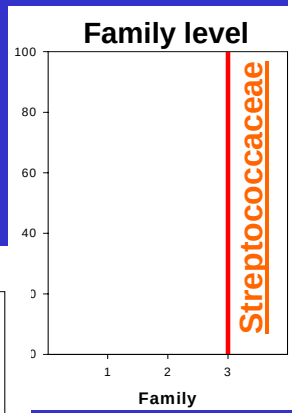
Identification of Bacteria Using Phylogenetic Relationships Revealed by MS/MS Sequencing of Tryptic Peptides

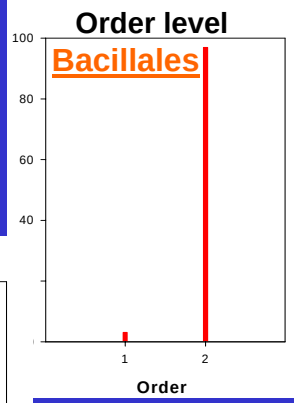
(*Lactococcus lactis*)

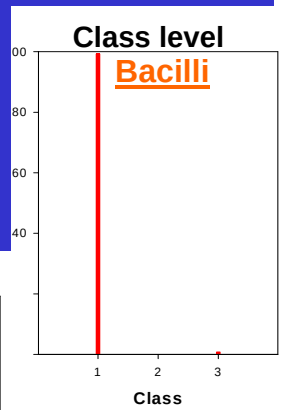


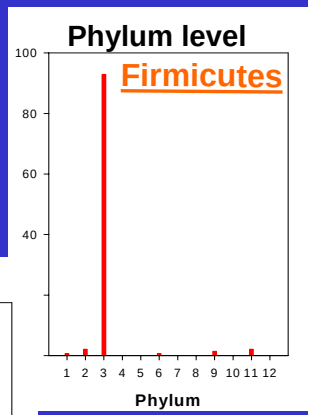
F_2

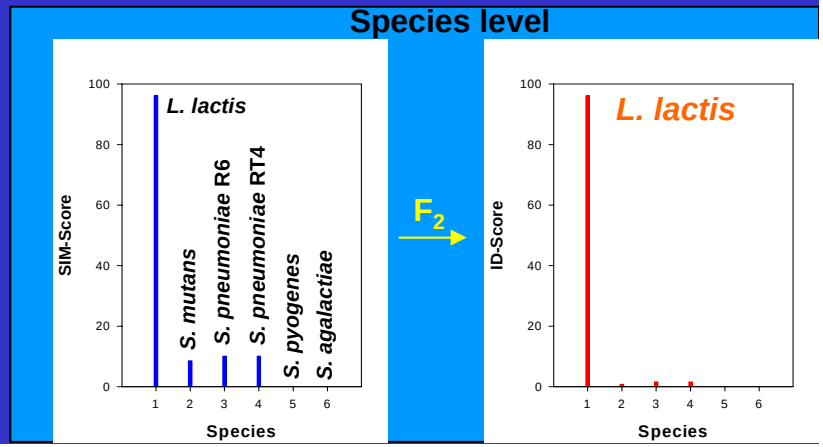








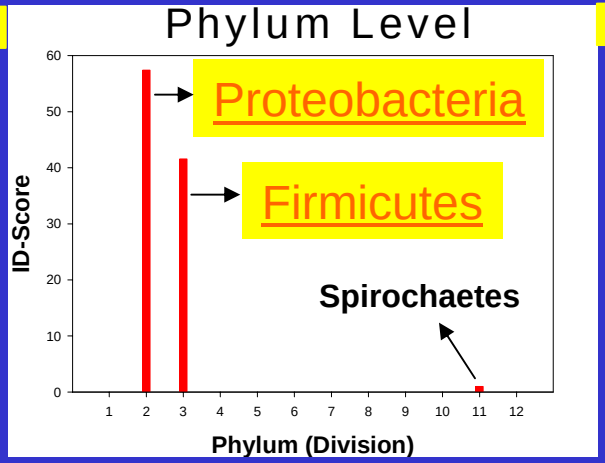




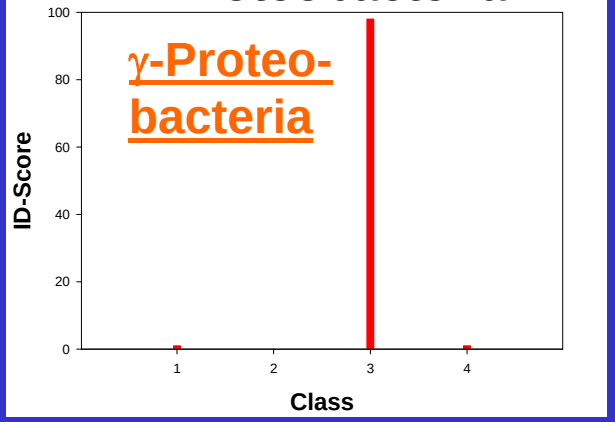
Analysis of Bacterial Mixture

*[E. coli (*K-12) and *B. subtilis* (2:1), w:w]

Phylum Level



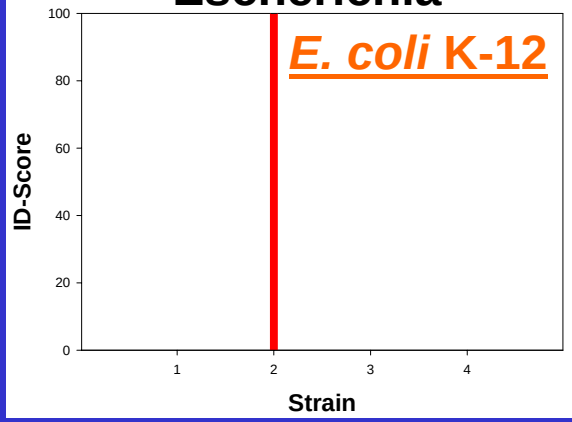
Proteobacteria



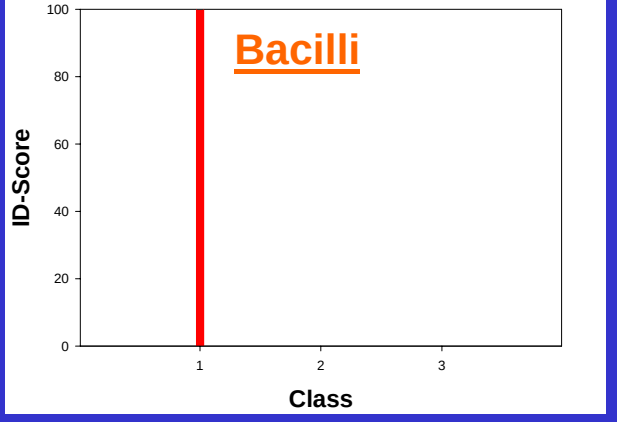
Enterobacteriales

Enterobacteriaceae

Escherichia



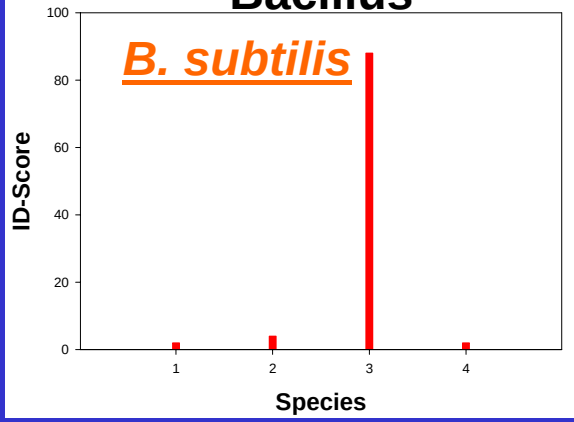
Firmicutes



Bacillales

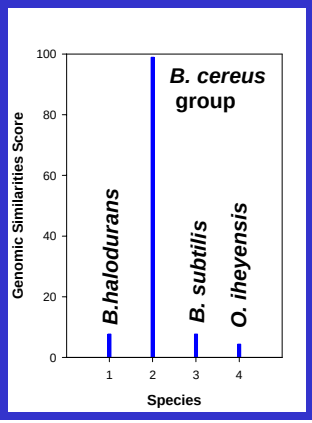
Bacillaceae

Bacillus

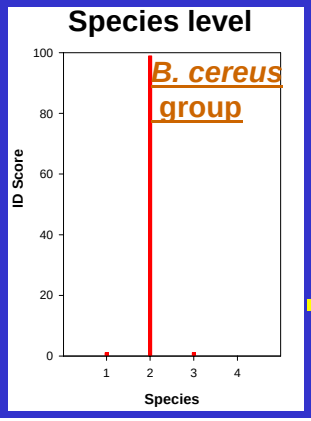


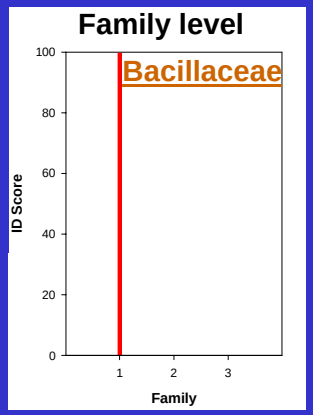
Identification of Bacteria Using Phylogenetic Relationships Revealed by MS/MS Sequencing of Tryptic Peptides

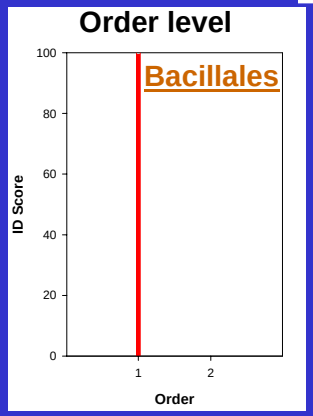
(*Bacillus cereus*)

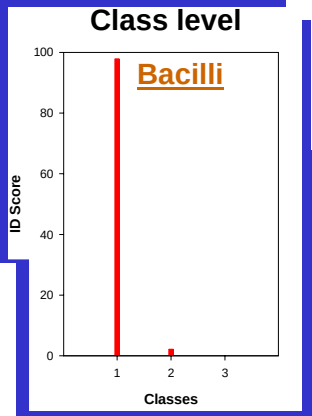


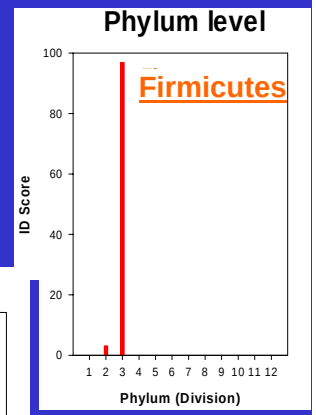
F_2

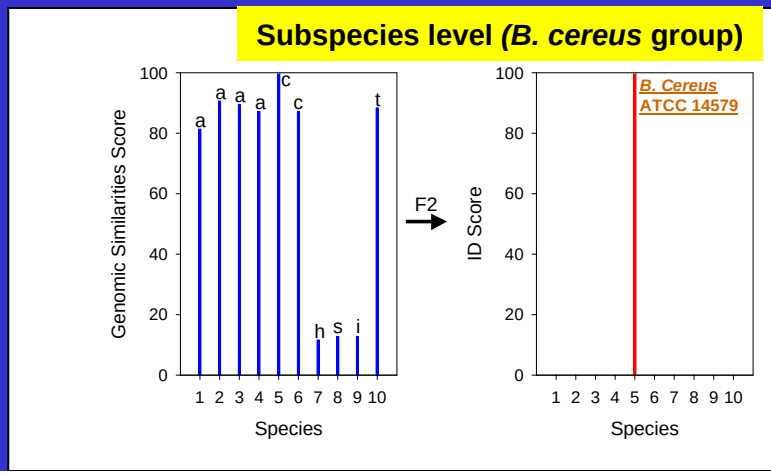




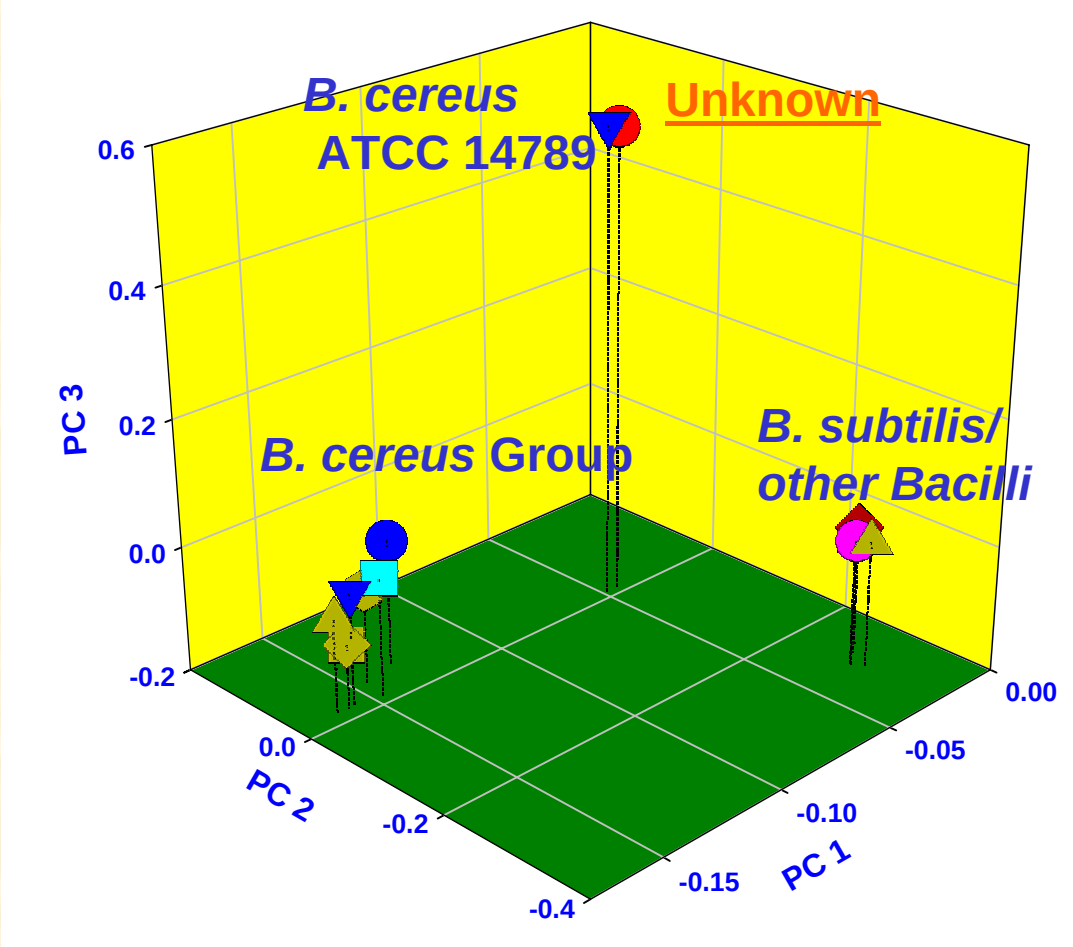






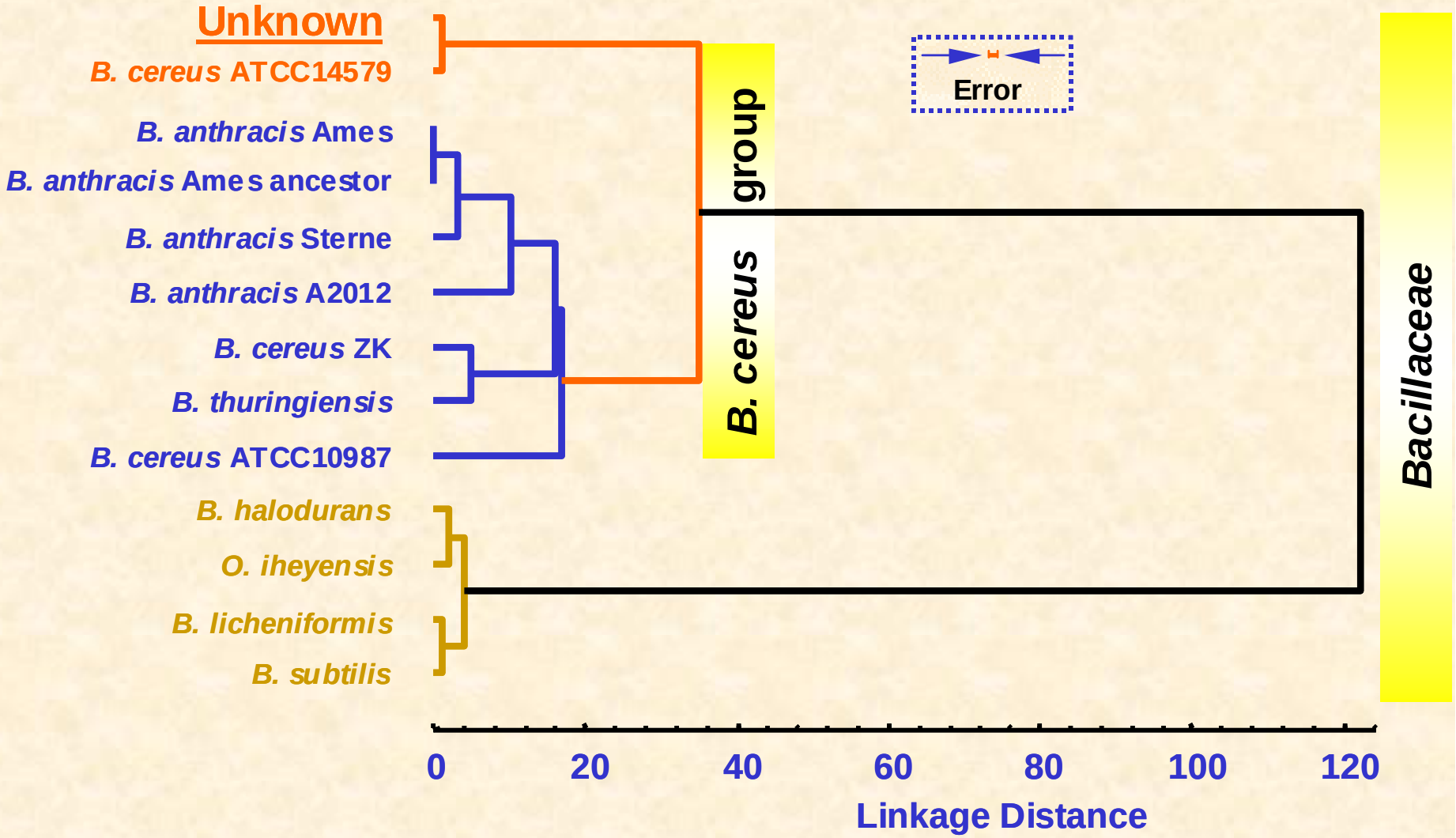


Principal Component Analysis of Peptide Assignments



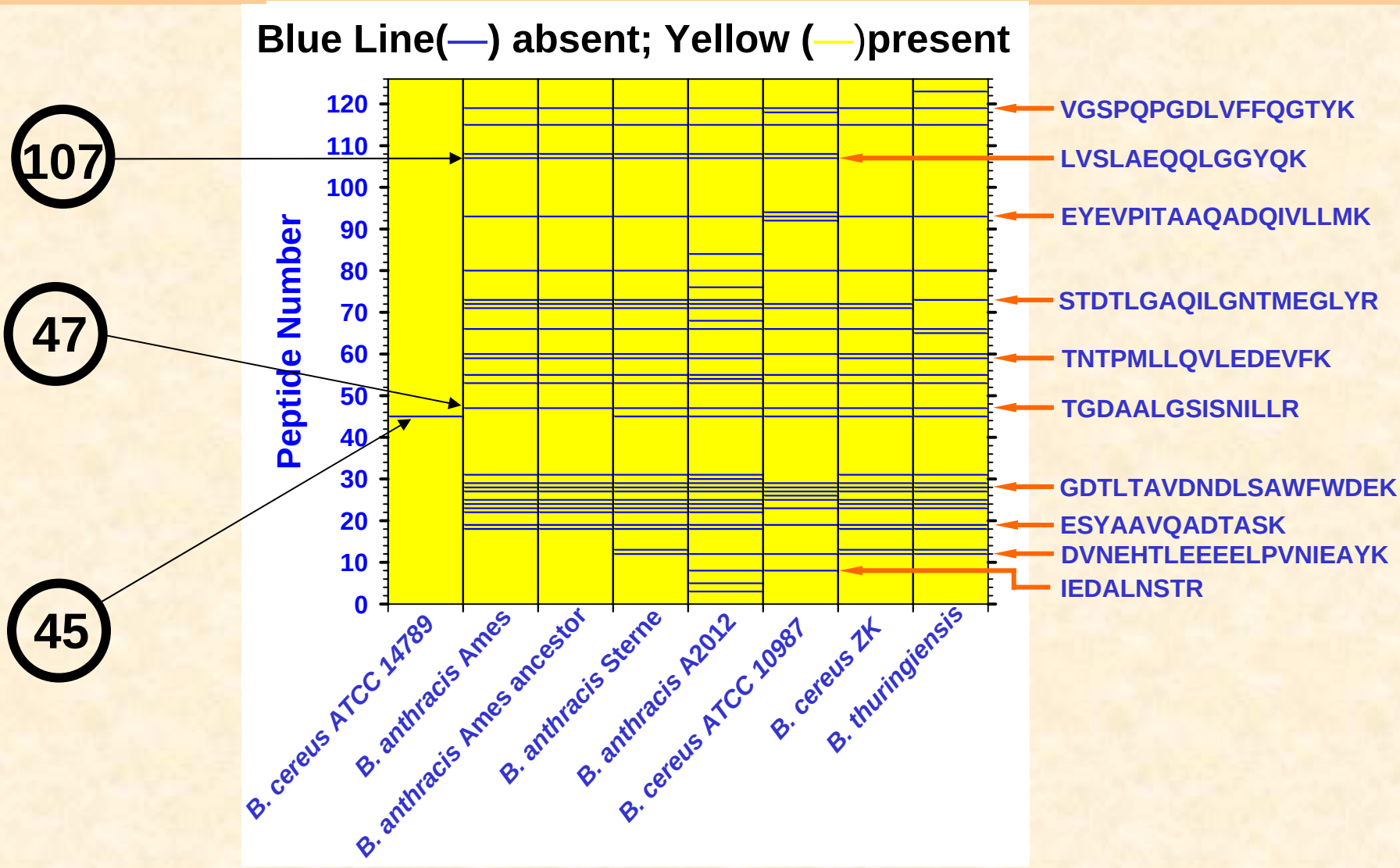
Representation of the database *Bacillaceae* species and unknown organism in the principal component space (PC 1, PC 2, PC 3) reflecting 80 % of the total information included in the assignment matrix of 125 amino acid peptide sequences to bacterial proteomes.

Cluster Analysis of Peptide Assignments



Hierarchical clustering of *Bacillaceae* species in 125-dimensional space of peptide sequences.
(Complete linkage; squared Euclidean distances)

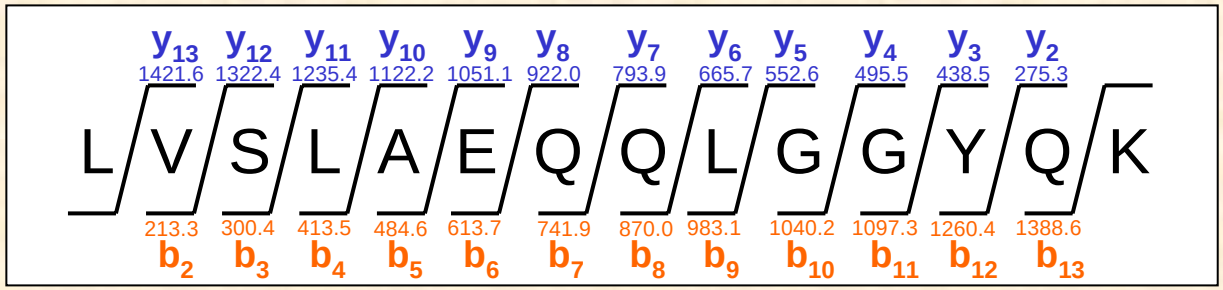
Differences in Proteome Composition Between an Unknown Sample and Database Bacteria

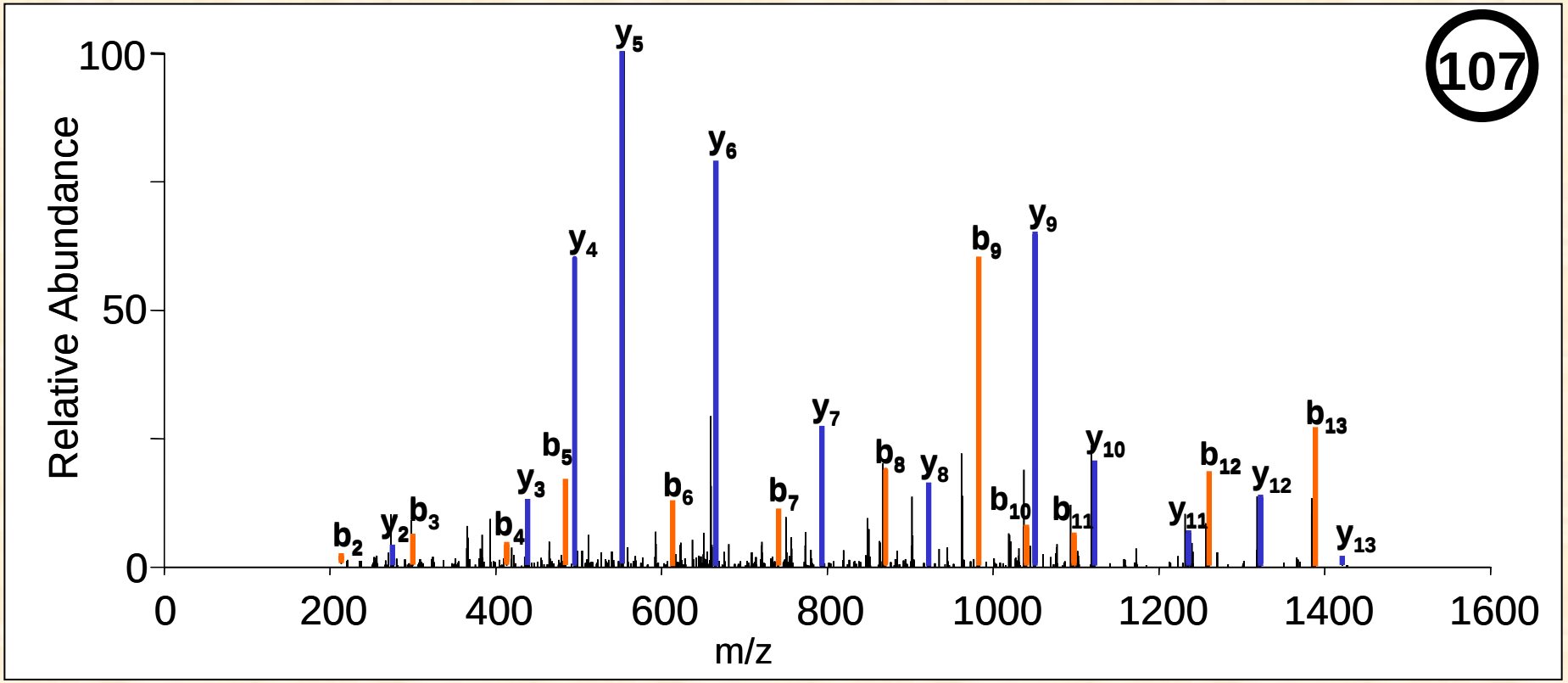


Unknown sample: correctly identified as *B. cereus* ATCC 14789

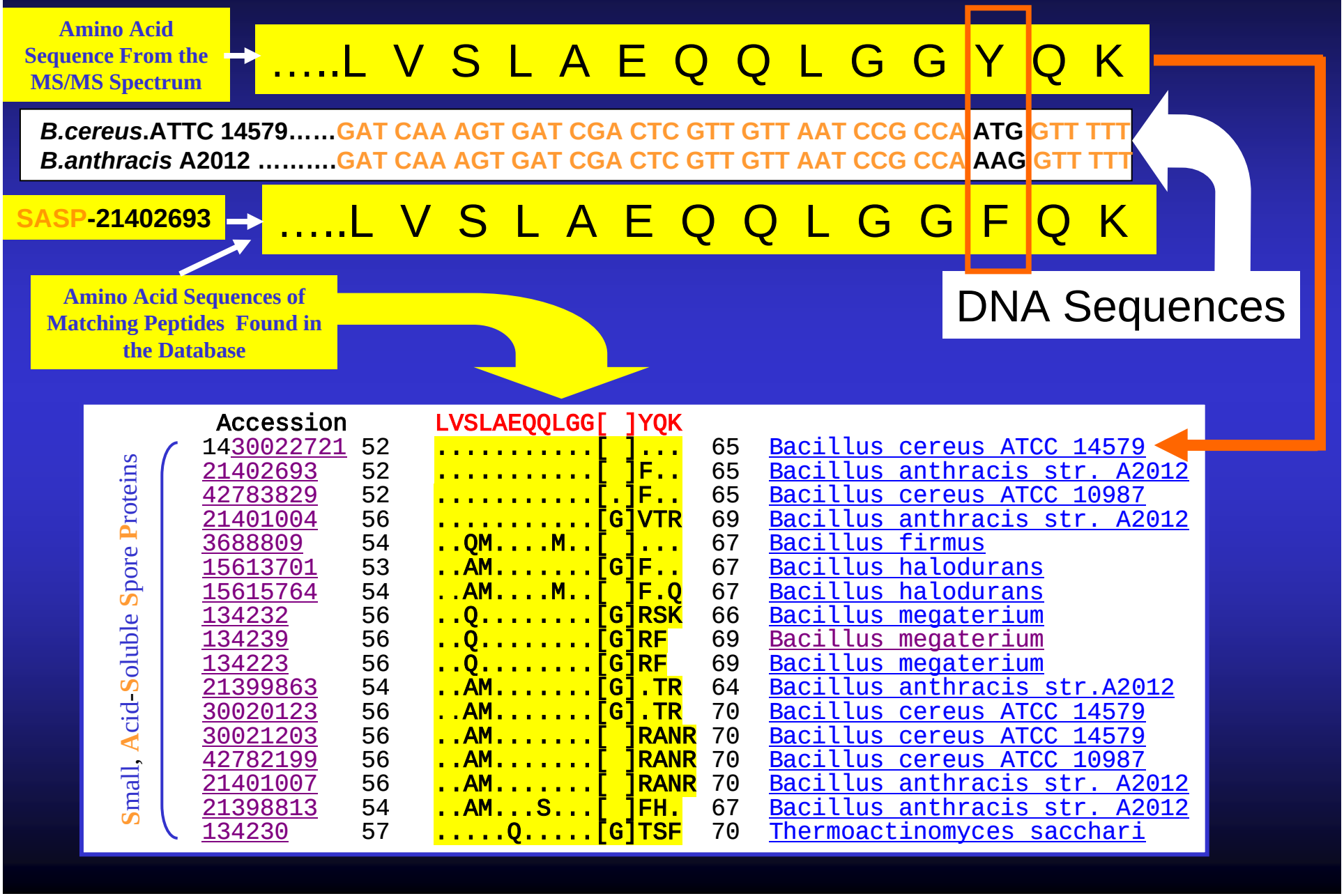
Product Ion Mass Spectrum of a Peptide Ion

Amino Acid Sequence Information Obtained in Less than 1 second





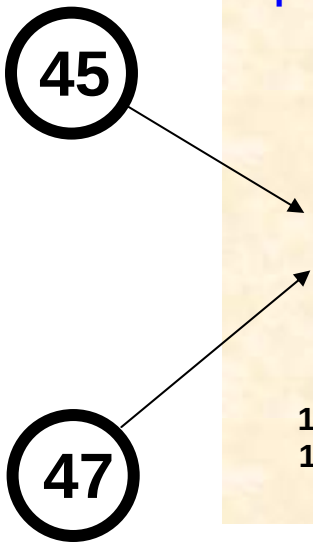
Discriminative Power of DNA and Protein Sequences



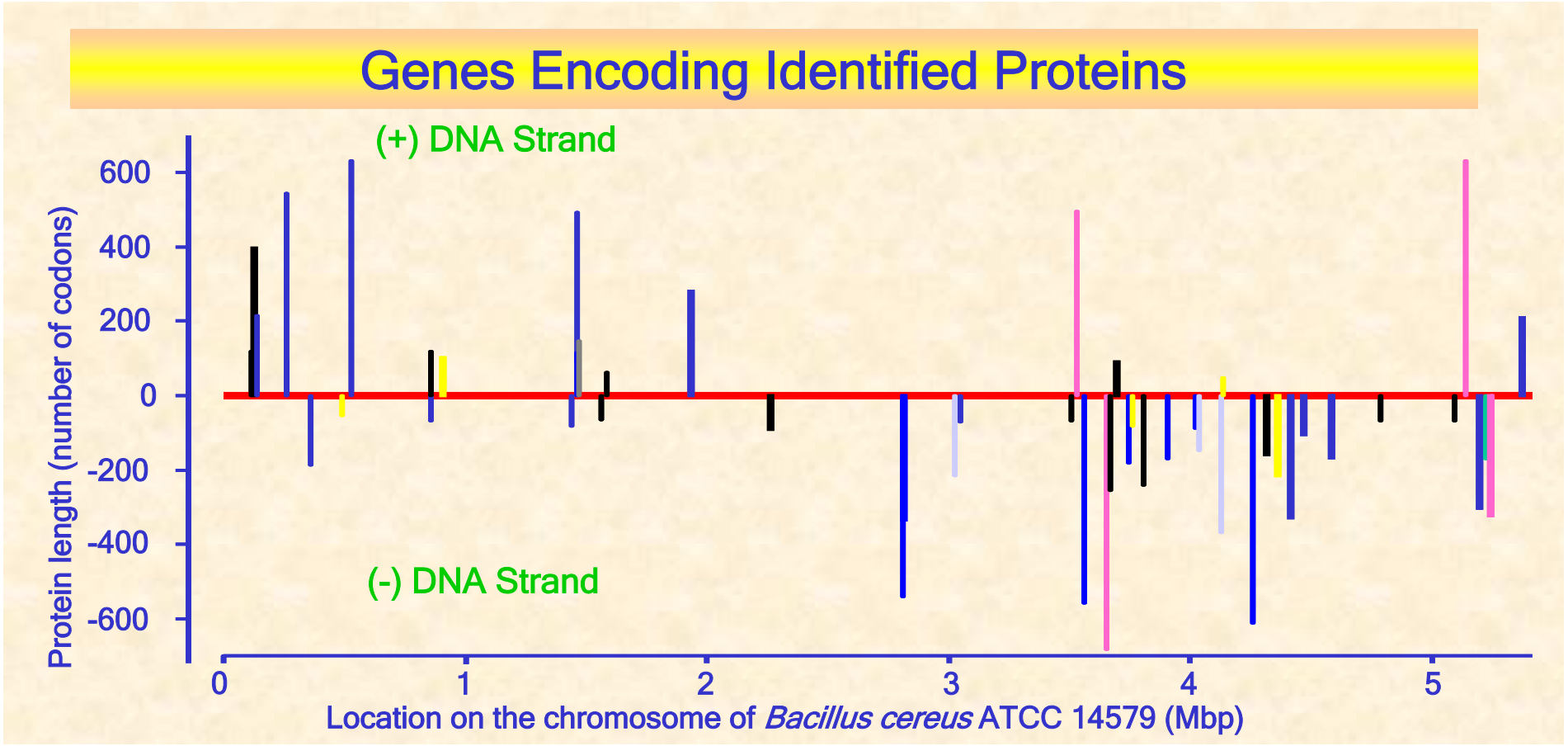
Small, Acid-Soluble Spore Proteins

Accession		LVSLAEQQLGG[ ]YQK		
1430022721	52	..........[ ]...	65	Bacillus cereus ATCC 14579
21402693	52	..........[ ]F..	65	Bacillus anthracis str. A2012
42783829	52	..........[.]F..	65	Bacillus cereus ATCC 10987
21401004	56	..........[G]VTR	69	Bacillus anthracis str. A2012
3688809	54	..QM....M..[ ]...	67	Bacillus firmus
15613701	53	..AM.......[G]F..	67	Bacillus halodurans
15615764	54	..AM....M..[ ]F.Q	67	Bacillus halodurans
134232	56	..Q........[G]RSK	66	Bacillus megaterium
134239	56	..Q........[G]RF	69	Bacillus megaterium
134223	56	..Q........[G]RF	69	Bacillus megaterium
21399863	54	..AM.......[G].TR	64	Bacillus anthracis str.A2012
30020123	56	..AM.......[G].TR	70	Bacillus cereus ATCC 14579
30021203	56	..AM.......[ ]RANR	70	Bacillus cereus ATCC 14579
42782199	56	..AM.......[ ]RANR	70	Bacillus cereus ATCC 10987
21401007	56	..AM.......[ ]RANR	70	Bacillus anthracis str. A2012
21398813	54	..AM...S...[ ]FH.	67	Bacillus anthracis str. A2012
134230	57	.....Q.....[G]TSF	70	Thermoactinomyces sacchari

Selected Peptide Sequences Discriminating Between an Unknown and Database Bacteria



Peptide #	Sequence	Protein
8	IEDALNSTR	60 kDa chaperonin GROEL
12	DVNEHTLEEEELPVNIEAYK	Hypothetical protein BC0479
19	ESYAAVQADTASK	Putative transcriptional regulator
29	GDTLTAVDNDLSAWFWDEK	Spore coat-associated protein N
45	KQPNFDDSSNFAK	Hypothetical protein BA 3347 [*Bacillus anthracis* Ames]
47	TGDAALGSISNILLR	Flagellin
59	TNTPMLLQVLEDEVFK	Propionyl-CoA carboxylase biotin-containing subunit
73	STDTLGAQILGNTMEGLYR	Oligopeptide-binding protein oppA
93	EYEVPITAAQADQIVLLMK	IG hypothetical 17696
107	LVSLAEQQLGGYQK	Small acid-soluble spore protein
119	VGSPQPGDLVFFQGTYK	N-acetylmuramoyl-L-alanine amidase


Genes Encoding Identified Proteins
(+) DNA Strand
(-) DNA Strand
Protein length (number of codons)
600
400
200
0
-200
-400
-600
0
1
2
3
4
5
Location on the chromosome of *Bacillus cereus* ATCC 14579 (Mbp)

CONCLUSIONS

The results demonstrate that mass spectrometry-based proteomics approach allows for :

- High confidence level classification and identification of bacteria based on genome traceable, proteomic similarities and differences between an analyzed microorganism and reference bacteria;

- Identification of pure cultures as well as mixtures of microorganisms.


Thank you !!!